

КРАТКИЕ СООБЩЕНИЯ

УДК 519.6: 543.545

© И. А. Леонтьев

ОБСЧЕТ ПИКОВ В ЗАДАЧАХ ЭЛЕКТРОФОРЕЗА

Электрофоретические задачи требуют обработки данных. В работе рассматриваются методы для обсчета данных в задачах электрофореза ДНК. Эти методы могут применяться и для многих других задач. Работа посвящена в основном методам просчета параметров гауссовых пиков. Это может помочь в количественном анализе.

ВВЕДЕНИЕ

Информационные сигналы, полученные с детекторов прибора, как правило, подлежат первичной и вторичной обработке. Первичная обработка — выделение сигнала из шума [1]. Вторичная обработка — определение параметров сигналов [1]. В статье описаны методы вторичной обработки применительно для задач электрофореза ДНК. После первичной обработки сигнал представляет собой набор пиков, отличающихся друг от друга в основном временем выхода. Для определения местоположения пиков применяются методы поиска экстремальных точек — максимумов. Для их поиска применяются различные методы, описанные в литературе. Один из методов предложен в статье [2].

После того как положение максимумов, а значит, и примерное положение пиков известно, необходимо определить их параметры. В качестве модели пика была выбрана гауссова кривая, т. е. $f(t) = x_1 \cdot e^{x_2(t-x_3)^2}$. Ее параметрами являются амплитуда (x_1), ширина $\sigma = \sqrt{\frac{1}{2x_2}}$ и положение (x_3).

В статье [2] был предложен метод для определения амплитуды, ширины и положения пиков с помощью квазиньютоновских методов. Здесь будет приведен результат работы данного метода.

ОПРЕДЕЛЕНИЕ ПАРАМЕТРОВ ПИКА

Метод, описанный в [2], состоит в следующем. Если $f(t) = x_1 \cdot e^{x_2(t-x_3)^2}$ и имеется набор измерений f_i , то задача сводится к поиску

$$\min_{x_1, x_2, x_3} F(x_1, x_2, x_3) = \sum_{i=1}^m (f_i - x_1 \cdot e^{x_2(t_i - x_3)^2})^2. \quad (1)$$

Эта задача носит название нелинейной задачи наименьших квадратов [3]. Задача оптимизации

состоит в минимизации $F(x)$. Алгоритмы для минимизации функции n переменных разрабатываются уже свыше 140 лет. Наиболее распространенными являются метод наискорейшего спуска и многомерный метод Ньютона. Достоинство метода наискорейшего спуска заключается в том, что он всегда сходится, если функция $F(x)$ имеет непрерывные производные. Но в некоторых случаях он сходится довольно медленно (иногда требуется более ста итераций).

Этого можно избежать, применяя n -мерный аналог метода Ньютона. Функция F записывается в виде

$$\begin{aligned} F(x) &\approx \\ &\approx F(x) + p^T \nabla F(x) + \frac{1}{2} p^T \nabla^2 F(x) p = \\ &= F(x) + Q(p). \end{aligned} \quad (2)$$

Чтобы получить шаг p , минимизируется квадратичная функция $Q(p)$ при помощи построения ее градиента по p . В конечном итоге получается:

$$x_{k+1} = x_k + p = x_k - \nabla^2 F(x_k)^{-1} \nabla F(x_k). \quad (3)$$

Метод Ньютона в n -мерном случае обладает свойством быстрой сходимости, а именно он сходится квадратично в окрестности решения

$$\|x_{k+1} - x^*\|_n \leq \beta \|x_k - x^*\|_n^2, \quad (4)$$

где β — некоторая неотрицательная константа, зависящая от $F(x)$.

Но у метода Ньютона также есть недостатки. Например, он может не сходиться. Но если матрица $\nabla^2 F^{-1}$ положительно определена, т. е. удовлетворяет условию $z^T \nabla^2 F^{-1} z > 0$ для всех $z \neq 0$, то в этом случае направление метода Ньютона гарантировано будет направлением спуска [3].

Еще одним недостатком метода Ньютона является необходимость вычислять матрицу вторых производных [3].

Для того чтобы получить альтернативные эффективные и практичные методы, можно аппроксимировать гессиан в ходе минимизации функции. Эти методы основаны на аппроксимации гессиана текущими и являются обобщением метода текущих. Если $B_k \approx \nabla^2 F_k = F(x)$, то шаг на k -й итерации будет определяться из системы

$$B_k p = -\nabla F_k. \quad (5)$$

Таким образом, мы получаем вариант метода Ньютона, в котором используется приближенный гессиан. Новая аппроксимация гессиана выбирается так, чтобы

$$B_{k+1}(x_{k+1} - x_k) = \nabla F_{k+1} - \nabla F_k. \quad (6)$$

В одномерном случае это однозначно определяет B_{k+1} . При более высоких размерностях для того, чтобы определить B_{k+1} , необходимы дополнительные условия.

Поскольку нам необходимо на каждой итерации вычислять обратную матрицу B^{-1} , можно еще упростить метод Ньютона, используя алгоритм Шермана—Моррисона для приближенного вычисления обратной матрицы. Предположим, что для матрицы B_k вычислена обратная B_k^{-1} . Пусть на следующей итерации матрица

$$B_{k+1} = B_k - uv^T, \quad (7)$$

где u и v — векторы размерности n . Тогда матрицу B_{k+1}^{-1} можно вычислять по формуле

$$B_{k+1}^{-1} = B_k^{-1} + \alpha (B_k^{-1}u)(v^T B_k^{-1}), \quad (8)$$

где $\alpha = 1/(1 - v^T B_k^{-1}u)$. На это будет затрачено $O(n^2)$ арифметических операций по сравнению с $O(n^3)$ операций при стандартном вычислении новой обратной матрицы.

Для упрощения выбора векторов u и v полагаем, что $u_i = v_i$. Это обосновано, поскольку матрица вторых производных является симметричной. Для получения более достоверного результата элемент $\{B_{11}\}_k$ вычислялся как $d^2 F/dx^2$.

Важно, что достаточно хорошо нужно подобрать центр пика. Для этого применялся метод поиска центра масс. Для уточнения полученного результата использовался метод наименьших квадратов отдельно для параметра x_3 . И только после этого можно применять описанный выше метод.

Провести оценку метода можно по стандартному отклонению, стандартным ошибкам, доверительным интервалам или по величине R -квадрат. Но большинство этих оценок предназначено для оценки состоятельности модели и применимости метода наименьших квадратов. Чтобы оценить работу самого метода, здесь использовался график шкалированных невязок, т. е. невязок, деленных

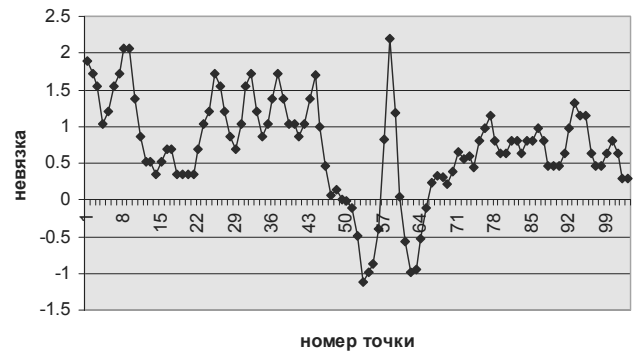


Рис. 1. График шкалированных невязок для предложенного метода

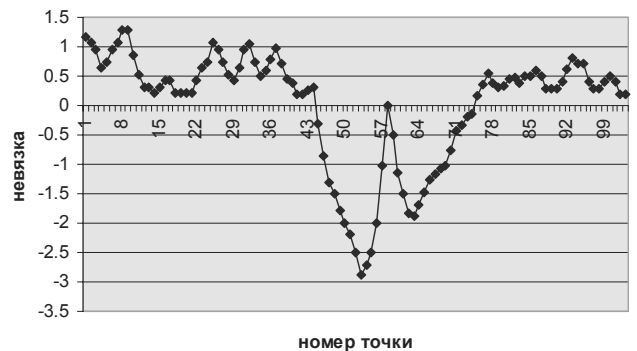


Рис. 2. График шкалированных невязок для метода сравнения

на стандартное отклонение. Примерно 95 % шкалированных невязок должно лежать в интервале $[-2, 2]$ [3].

График невязок, показанный на рис. 1, был взят для случайного пика. Пик аппроксимировался по 103 точкам. Из них 3 значения не попали в интервал $[-2, 2]$, что составляет примерно 3 %.

Для сравнения результатов параметры данного пика были вычислены другим, более простым методом. В этом методе: амплитуда пика — это его максимальное значение; полуширина — вычислена по его площади ($S = \sqrt{2\pi\sigma h}$); положение — вычислено методом наименьших квадратов. На рис. 2 приведен график шкалированных невязок для этих вычислений. Из графика видно, что 6 точек не лежат в интервале $[-2, 2]$, что составляет примерно 6 % и, следовательно, является неудовлетворительным результатом.

Предложенный метод достаточно прост в реализации и не требует большого объема вычислений. В результате использования данного метода

получены результаты, подтверждающие необходимость его использования.

ЗАКЛЮЧЕНИЕ

В статье приведены результаты испытания метода для обсчета пиков в электрофорезе ДНК. Метод позволяет достаточно быстро определить параметры пиков. График невязок показал, что метод достаточно хорошо работает для таких задач. Однако большинство невязок лежит в положительной области, что говорит о неидеальности метода и выбранной модели для описания пиков.

СПИСОК ЛИТЕРАТУРЫ

1. Сизиков В.С. Математические методы обработки результатов измерений. СПб.: Политехника, 2001. 101 с.

2. Леонтьев И.А. Обработка данных в задачах электрофореза // Научное приборостроение. 2003. Т. 13, № 2. С. 96–99.
3. Каханер Д., Моулер К., Нэш С. Численные методы и программное обеспечение. М.: Мир, 2001. С. 238–339, 408, 409, 424–429.

*Институт аналитического приборостроения РАН,
Санкт-Петербург*

Материал поступил в редакцию 13.01.2004.

PEAK CALCULATIONS IN ELECTROPHORESIS PROBLEMS

I. A. Leontyev

Institute for Analytical Instrumentation RAS, Saint-Petersburg

The paper is devoted to data processing methods as applied to DNA electrophoresis. These methods can be also used for many other problems. The main attention is given to the calculation of some characteristics of Gaussian peaks. This may be helpful in quantitative analysis.